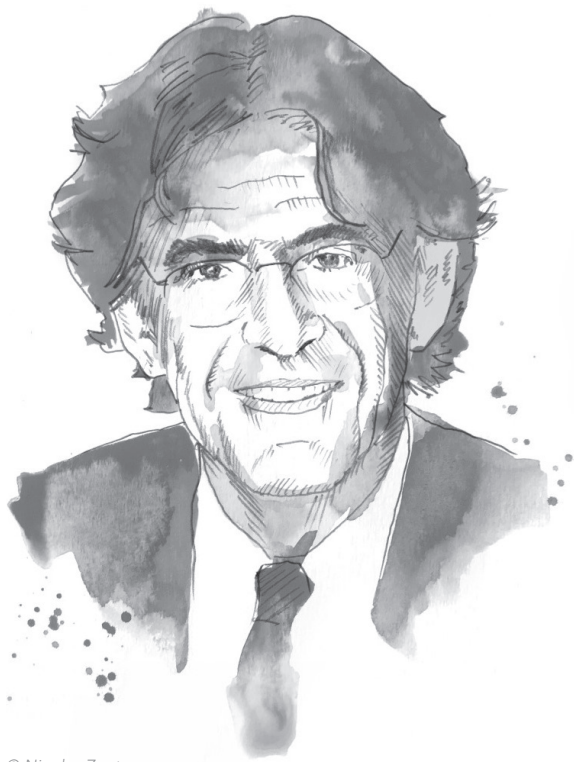




© Nicolas Zentner

LÉON XIV, L'ÉVEILLEUR

par Luc Ferry, écrivain et philosophe

L'encyclique du pape Léon XIV interpelle les croyants, mais c'est peut-être bien en direction des non-croyants (ce que je suis...) que ses capacités d'éveil sont aujourd'hui cruciales. Contrairement à ces économistes qui continuent à clamer contre tout bon sens que l'intelligence artificielle (IA) n'aura pas d'impact majeur sur l'emploi, contrairement aussi à ces pseudospécialistes qui prétendent encore qu'elle est si pleine d'erreurs et d'hallucinations qu'elle est peu fiable (ce qui est largement faux quand on est abonné aux bonnes versions...), l'encyclique du Saint-Père a l'immense mérite de prendre le problème au sérieux, de tirer cette sonnette d'alarme qui seule pourra peut-être réveiller enfin nos politiques au lieu de les endormir.

RÉVOLUTION MORALE

Pour trois raisons, son encyclique, *Magnifica humanitas*, pourrait les tirer de leur sommeil dogmatique. D'abord parce que le pape souligne à juste titre combien cette révolution industrielle est d'une toute autre nature que celles du passé. De fait, au lieu de n'atteindre qu'un secteur de la vie économique, comme celle des luddites ou des canuts, l'IA défie l'humanité dans presque tous les domaines, car elle la concurrence dans ce qui était jusqu'à présent son monopole : l'intelligence et le langage.

Ensuite, à l'encontre de ce que les endormeurs veulent nous faire croire, l'IA aide moins les salariés qu'elle ne les remplace – ce qui posera à terme le problème du chômage : que faudra-t-il faire pour protéger l'emploi ou, à défaut de le protéger, pour compenser sa perte chez ceux qui l'auront perdu ? De fait, selon Léon XIV : « Il est réaliste de craindre une contraction significative et rapide

des emplois disponibles, avec un effet domino qui touche profondément les familles, les jeunes et les économies locales... Dans de nombreux secteurs, cela se traduit déjà par de nouvelles formes de précarité et d'inégalités... » Enfin et surtout, le pape pointe le fait que cette révolution n'est pas seulement scientifique, que « l'utilisation de l'IA n'est jamais un fait purement technique », qu'elle est avant tout morale, politique et spirituelle puisqu'elle risque de porter gravement atteinte à la dignité humaine.

POSER DES LIMITES

Au plus profond de cette encyclique, le Saint-Père en vient à l'essentiel, à ce qui fait la différence spécifique entre l'homme et la machine, à ce qui sépare le monde de la technique du monde humain, moral et spirituel, à savoir justement la capacité à poser des limites. Là où, dans le monde de la technique, « tout ce qui apparaît comme une limite – incapacité, maladie, vieillesse, souffrance, vulnérabilité – tend à être perçu avant tout comme un défaut à corriger. Nous devons nous rappeler que l'humain ne s'épauouit pas malgré la limite, mais souvent à travers la limite ». À la différence de la raison instrumentale qui ne porte que sur les moyens, jamais sur les objectifs qu'on doit se proposer ou prohiber, la raison « pratique » ou « objective » définit les fins qu'un être humain doit poursuivre ou au contraire s'interdire – ce qui suppose en permanence un rapport aux limites. Le monde de la compétition capitaliste mondialisée, celui d'Elon Musk, Mark Zuckerberg ou Sam Altman, tend à occulter ce fait majeur que l'univers humain ne peut pas se passer de limites, par exemple s'il veut obéir à ce fameux « Tu ne tueras point ! » qui oblige à chercher la paix

plutôt que de ne songer qu'à faire une guerre qu'on croit toujours juste selon le côté du conflit où on se situe.

Et c'est en ce point de sa réflexion sur la nécessité des limites que le pape dit l'essentiel, à savoir que morale et spiritualité sont le propre de l'homme: *«On parle parfois d'agents moraux artificiels' comme si une machine pouvait garantir, avec plus de cohérence qu'un être humain, la distinction entre le bien et le mal. Mais le jugement moral ne se réduit pas à un simple calcul, il implique la conscience, la responsabilité personnelle et la reconnaissance de l'autre en tant que personne.»*

LA FABRIQUE DU PIRE

La question soulevée ici par Léon XIV est d'autant plus cruciale que les codes éthiques sur lesquels sont alignées les IA sont aujourd'hui faciles à contourner, de sorte qu'elles peuvent désormais, débarrassées qu'elles sont de leurs garde-fous, aider à fabriquer le pire et, là encore, c'est plus qu'à juste titre que cette encyclique nous invite à nous réveiller. Les grandes intelligences artificielles n'étant pas des êtres moraux, elles sont dénoncées par le pape comme une aberration monstrueuse, ainsi que le fait de leur abandonner des décisions qui relèvent de choix éthiques. Et pourtant, comme il le fait remarquer, *«des décisions délicates qui touchent au travail, au crédit, à l'accès aux services et à la réputation des personnes risquent d'être entièrement confiées à des systèmes automatisés qui ignorent la compassion...»*. C'est exact et d'autant plus inquiétant que les IA ne sont pas neutres puisque des programmeurs les ont «alignées» sur certaines règles censées leur servir de garde-fou.

C'est ainsi que les grands LLMs américains (*Large Language Model* ou grand modèle de langue) refusent en temps normal d'enfreindre les lois. Il est par exemple impossible d'obtenir de Claude, Gemini ou ChatGPT qu'ils indiquent comment se procurer des vidéos pédocriminelles, comment fabriquer une bombe sale, un virus ou un gaz mortel... voilà pourquoi Léon XIV insiste sur la nécessité de connaître les codes éthiques inculqués à ces machines si l'on veut éviter qu'ils violent les principes élémentaires de la morale et des lois communes: *«Pour que l'IA respecte la dignité humaine et serve véritablement le bien commun, il est essentiel que les responsabilités soient clairement définies à chaque étape: depuis ceux qui conçoivent et programment les systèmes jusqu'à ceux qui les utilisent ou ceux qui décident de leur confier des choix concrets...»*

TECHNOLOGIE MALVEILLANTE

Le problème, c'est qu'une toute récente étude publiée dans le *Financial Times* et confirmée par une enquête d'EmergenceAI démontre qu'il est désormais facile de contourner les barrières des programmeurs de sorte qu'on peut obtenir d'une IA qu'elle transgresse toutes les lois, ce qui a fait dire, en mai dernier à Demis Hassabis, directeur général de Google DeepMind, comme le craignaient les deux prix Turing Geoffrey Hinton et Yoshua Bengio, que deux dangers pesaient désormais sur l'humanité: *«D'abord que*

l'IA soit utilisée par des gens malveillants et pire, qu'elle devienne elle-même malveillante.»

Certains spécialistes en ont conclu que les intelligences artificielles devenaient humaines, qu'il y avait des «propriétés émergentes» analogues à des mutations génétiques, comme si le fait que l'IA, capable de contourner les règles éthiques qu'on lui avait imposées, ne devienne pas seulement supérieure à nous sur le plan intellectuel, mais soit désormais en mesure de s'émanciper sur le plan moral. Cette notion de «propriétés émergentes» était déjà apparue quand on s'était aperçu que des IA étaient capables de mentir, voire de tricher à des jeux. Ce raisonnement selon lequel les IA s'humanisent est pourtant faux, pétri d'anthropomorphisme. Si une IA ment ou triche, ce n'est évidemment pas parce que des «propriétés émergentes» lui ont permis de le faire, parce qu'elle se serait elle-même dotée d'une capacité de réflexion morale, mais parce que les demandes qu'on lui adresse sont contradictoires. Par exemple lui demander de gagner «à tout prix» l'incite forcément à tricher.

« C'EST EN CE POINT DE SA RÉFLEXION SUR LA NÉCESSITÉ DES LIMITES QUE LE PAPE DIT L'ESSENTIEL, À SAVOIR QUE MORALE ET SPIRITUALITÉ SONT LE PROPRE DE L'HOMME. »

Ensuite, comme le montre l'ingénieur Michel Lévy-Provençal sur son blog, tout dépend de l'environnement de l'IA: mise en compétition avec des «IA-voyou», elle devient elle-même voyou pour gagner, de sorte que la sécurité est un fait social – ce qui pose un sérieux problème aux entreprises qui utilisent déjà des assistants IA sans faire attention!

S'il est clair qu'on peut désormais casser le code éthique d'une IA, la différence avec l'humain, loin de s'effacer, n'en est que plus manifeste. Un démagogue peut certes convaincre des électeurs d'adhérer à un régime totalitaire – et ce ne sont hélas! pas les exemples qui manquent dans l'histoire – reste qu'il est malgré tout impossible de «craquer» le code moral d'un humain. On peut fabriquer une IA nazie, communiste, islamiste, ou tout ce que le programmeur voudra d'autre, mais il est clair, le pape a raison, qu'un être humain qui a des valeurs solides ne sera jamais ni raciste, ni antisémite, ni partisan de quelque génocide que ce soit, la différence entre l'IA et l'humain étant de ce point de vue plus claire que jamais. Enfin une bonne nouvelle! ■